文章编号: 1671-8844(2003)04-129-03

基于数据仓库和数据挖掘的教育决策支持系统

张德新[1], 崔 巍[2], 蒋天发[3]

(1. 郧阳师范高等专科学校物理系, 湖北 丹江口 442700; 2. 武汉大学遥感信息工程学院, 湖北 武汉 430079;
3. 中南民族大学计算机科学学院, 湖北 武汉 430074)

摘要: 在讨论了数据仓库和数据挖掘技术的基本概念、数据的组织方式和系统设计方法的基础上, 提出了基于数据仓库和数据挖掘的决策支持系统框架模型, 并将其应用到教育决策支持实践, 从而提高人才素质评价的高效性和科学性.

关键词: 数据仓库; 数据挖掘; 人才素质; 决策支持系统

中图分类号: TP 311 文献标识码: A

Education decision support system based on data warehouse and data mining

ZHANG De-xin[1], CUI Wei[2], JIANG Tian-fa[3]

(1. Department of Physics, Yunyang Teachers College, Danjiangkou 442700, China;
2. School of Remote Sensing Information Engineering, Wuhan University, Wuhan 430079, China;
3. Department of Computer Science, South-Central University for Nationalities, Wuhan 430074, China)

Abstract: On the basis of discussing the basic concepts of data warehouse and data mining technique, data organization scheme and method of system design; a model of decision support system framework is presented based on data warehouse and data mining. And its application to education decision support practice is carried out so as to heighten effectiveness and science of talent quality evaluation.

Key words: data warehouse; data mining; talent quality; decision support system

决策支持系统(decision support systems, 简写为 DSS)是在管理信息系统的基础上发展起来的[1], DSS 在实际应用开发过程中暴露出许多问题, 主要有以下两个方面: 一是 DSS 使用的数据库只能对原始数据进行一般的加工和汇总, 致使决策所需信息不足, 难以满足 DSS 的需要; 二是传统的 DSS 是模型驱动, 其模型库管理系统是 DSS 的核心部件, 但相对于决策本身的动态性和复杂性, 模型库提供的分析能力有限, 并且它所提供的模型独立于环境之外, 决策者和模型交互很少, 模型参数固定不变, 因此常常做出不正确的决策方案.

进入 20 世纪 90 年代后, 信息技术领域出现了数据仓库和数据挖掘技术的研究和开发热潮, 这为克服传统 DSS 存在的问题提供了技术上的支持, 为 DSS 开辟了一条新的途径. 目前开发的综合 DSS 是以数据仓库(data warehouse, 简写为 DW)技术为基础, 以数据挖掘(data mining, 简写为 DM)工具为手段进行实施的一整套解决方案.

1 基于数据仓库和数据挖掘的决策支持系统

1.1 数据仓库的基本概念

数据仓库是支持决策过程的、面向主题的、集

收稿日期: 2002-11-18
作者简介: 张德新(1955-), 男, 湖北郧西人, 副教授, 主要从事理论物理教学和数据库应用的研究.
基金来源: 湖北省教育厅重点科研资助项目(99S056)

成的、稳定的、不同时间的数据集合[2]. 具体分析如下:

(1)数据面向主题性. 主题是抽象的概念,对应客观分析领域的一个分析对象,根据决策确定主题后,就可围绕主题选择数据源. 基于主题组织的数据被划分为各自独立的领域,每个领域有自己的逻辑内涵,互不交叉. 例如,学生就是学校数据仓库的一个主题.

(2)数据集成与综合性. 为了反映用户单位内部复杂的全局模式,数据仓库既要集成用户单位内部各方面的数据,也要集成单位外的相关数据,为支持多层次多主题决策,数据仓库对数据要进行不同程度的综合.

(3)数据历史性. 数据历史性主要表现在两个方面:一是数据仓库内数据是关于各个主题不同时间的综合信息,多为 5 ~ 10 年;二是数据一旦进入数据仓库就不应更新,具有一定的稳定性.

数据多维组织模型(图 1)对决策影响最为关键. 由于决策所需的数据总是与维数(每一维代表对数据的一个特定的观察视角,如地区、时间等)和不同级别(如部门、单位、地区等)的统计和计算有关. 所以,决策支持系统以多维数据分析为核心,如用于人才素质评价的多维数据模型[3].

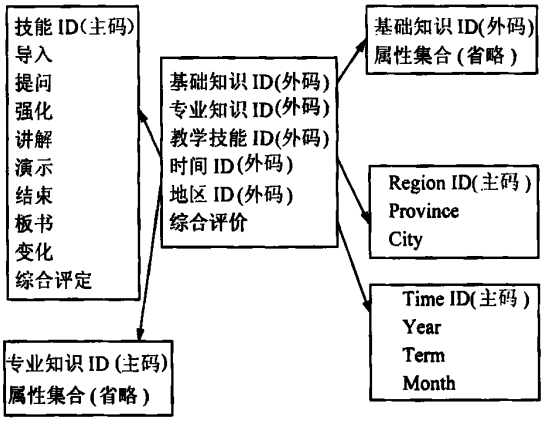

图 1 人才素质评价数据仓库的多维概念模型

1.2 数据挖掘的基本概念

数据挖掘是从大量数据中挖掘出隐含的、先前未知的、对决策有潜在价值的知识和规则[4]. 数据挖掘技术涉及数据库、人工智能、机器学习、神经计算和统计分析等多种技术,它使 DSS 跨入了一个新阶段. 数据挖掘的特点是:查询一般是决策制定者(用户)提出的即时随机查询,往往不能形成精确的查询要求;数据变化迅速,可能很快过时,因此需要对动态数据作出快速反应,以提供决策支持;主要基于大样本的统计规律,其发现的规则不一定适用于所有数据.

1.3 基于数据仓库和数据挖掘的决策支持系统

数据仓库和数据挖掘是作为两种独立的信息处理技术出现的,数据仓库技术用于数据的存储和组织,数据挖掘技术则致力于知识的自动发现. 由于这两种技术内在的联系性和互补性,为了充分发挥它们各自的特长,可以将它们结合起来,设计出一种新的 DSS 构架,即以数据仓库为基础、以数据挖掘工具为手段并以模型库为辅助决策的解决方案,其结构框图如图 2 所示.

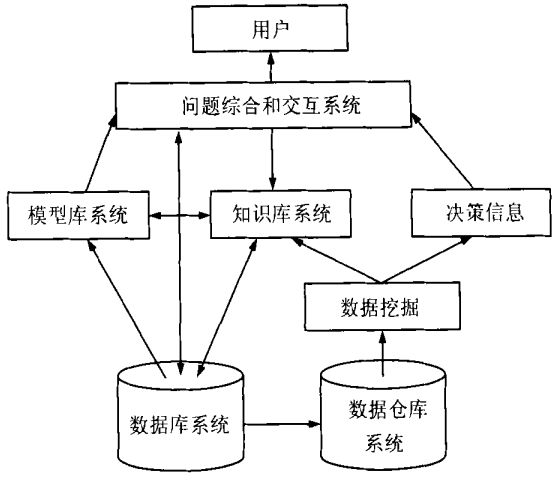

图 2 基于数据仓库和数据挖掘 DSS 构架图

基于数据仓库和数据挖掘 DSS 构架的主要特点是:数据仓库对底层数据库中的事务级数据进行集成、转换和综合,重新组织成面向全局的数据视图,为 DSS 提供数据存储和组织的基础. 数据挖掘以数据仓库的大量数据为基础,发现数据中的潜在模式,并以这些模式为基础做出预测. 数据挖掘表明,知识就隐藏在日常积累下来的大量数据之中,仅靠复杂的算法和推理并不能发现知识,数据才是知识的真正源泉. 在传统的 DSS 中,数据库、模型库和知识库往往被独立地设计和实现,因而缺乏内在的统一性,而数据仓库和数据挖掘组成的新的 DSS 构架解决了 DSS 数据库内数据不一致的问题. 由于内在的统一性,这种新结构很好地解决了相互间的衔接问题. 数据仓库为数据挖掘提供了充分可靠的数据基础,数据挖掘可以从数据仓库中找到所需的数据,挖掘出的知识可以直接用于指导决策分析处理过程并立即补充到系统的知识库中.

这种新的 DSS 构架真正展示了信息的本质, 表明了 DSS 的设计观念从模型驱动到数据驱动的转变.

2　用于人才素质评价的综合决策支持系统

基于数据仓库和数据挖掘的决策支持系统与传统的决策支持系统的本质区别在于, 新方法是在没有明确假设的前提下去挖掘信息、发现知识并由此做出决策. 数据挖掘所得到的信息应具有事先未知、有效性和实用性 3 个特征, 其中最重要的是事先未知. 决策过程的待求问题是数据挖掘过程的基础, 它驱动了整个数据挖掘过程, 也是检验最后结果和指引分析人员完成数据挖掘的依据. 图 3 描述了基于数据挖掘的决策基本过程. 决策过程中各步骤如下:

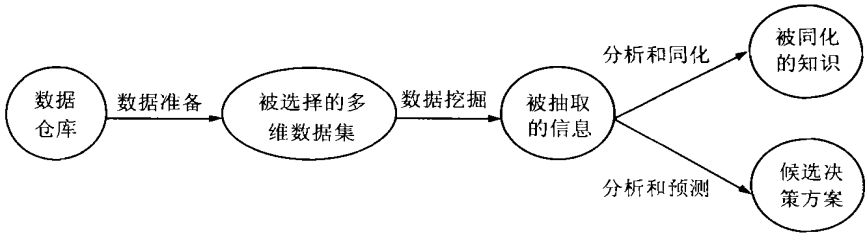

图 3　基于数据挖掘和数据仓库的决策过程

(1)确定待求解的问题. 清晰地定义出待求问题, 认清数据挖掘的目的是基于数据挖掘决策支持的重要一步. 数据挖掘的结论是不可预测的, 但要探索的问题应是有预见的.

(2)数据准备包括数据的选择、数据的预处理和数据的转换. 数据的选择是在数据仓库中搜索所有与问题有关的多维数据信息, 并从中选择出适用于数据挖掘应用的数据; 数据的预处理是研究数据的质量, 为进一步的分析做准备, 并确定将要进行的挖掘操作的类型; 数据的转换是将数据转换成一个分析模型, 这个分析模型是针对挖掘算法建立的. 建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键.

(3)数据挖掘, 对所得到的经过转换的数据进行挖掘, 除了完善选择合适的挖掘算法外, 其余一切工作都能自动完成. 新决策支持系统与传统决策支持系统的根本区别在于: 传统决策支持系统是根据选定的模型在数据库中准备数据的, 这样就无法真正地从实际问题的求解环境——数据出发, 做出的决策方案不可避免地会有一定的偏差甚至是完全错误; 而新系统完全是数据驱动的系统, 挖掘的知识和以此为基础产生的决策方案由数据确定, 事先未知. 如对人才进行分类的问题, 在旧系统中要事先选定一个分类模型, 然后对数据进行处理, 如果模型不适合数据, 就会产生错误的决策方案. 而新系统采用数据挖掘的聚类分析算法进行数据聚类, 产生的类的个数和类的特点是事先未知的, 这样完全由数据自身的特点决定分类的结果, 显然更合理和实用.

(4)分析和预测, 提出决策方案. 针对数据挖掘的初步结果, 利用分析和预测算法进行分析和预测, 再辅以基于模型库的推理, 即可产生合理的决策方案.

(5)分析和同化, 更新知识库. 将分析所得到的知识集成到系统的组织结构中去, 更新知识库内容.

3　结　论

由于新决策支持系统是以数据驱动为主, 解决了以模型驱动为主的传统决策支持系统做出决策方案往往出现偏差的缺陷, 做出的决策方案完全由数据自身(实际问题及其环境的数学抽象)的特点决定, 显然更合理和实用.

参考文献:

[1] 陈文伟 . 决策支持系统及其开发[M] . 第二版. 北京: 清华大学出版社, 2000.

[2] Inmon W H. 数据仓库[M] . 北京: 机械工业出版社, 2000.

[3] 张德新, 崔　巍. 数据仓库与人才素质评价[J]. 电脑与信息技术, 2002(1): 15.

[4] Jiawei Han . 数据挖掘: 概念与技术[M] . 北京: 机械工业出版社, 2001.